

Максимальная величина смещения эмпирического риска

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».
Лекция 9.

Аннотация

Будут рассмотрены вопросы определения точности оценок на основе эмпирического риска.

Основной вклад в погрешность даёт смещённость эмпирического риска. В асимптотике только она и остаётся. В частных случаях оказывается возможным сделать точные аналитические оценки.

Доверительный интервал в схеме Бернулли

Односторонний интервал $[0, \hat{p}]$

$$\sum_{i=0}^M C_N^i \hat{p}^i \cdot (1 - \hat{p})^{N-i} = \alpha.$$

Двусторонний интервал $[p_1, p_2]$

$$\sum_{i=0}^M C_N^i p_2^i \cdot (1 - p_2)^{N-i} = \sum_{i=M}^N C_N^i p_1^i \cdot (1 - p_1)^{N-i} = \frac{\alpha}{2}.$$

Нормальное приближение

Обозначим $\nu = \frac{M}{N}$.

Имеет место оценка

$$P(|\nu - p| > \varepsilon) \approx 1 - \Phi\left(\frac{\varepsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \leq 1 - \Phi(2\varepsilon\sqrt{N}) \leq e^{-2\varepsilon^2 N},$$

где $\Phi(x) = 2 \int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ – функция Лапласа.

Уточнение

Другой вариант

$$\mathbf{P} \left(\frac{|\nu - p|}{\sqrt{p(1-p)}} > 2\varepsilon \right) \approx 1 - \Phi(2\varepsilon\sqrt{N}) \leq e^{-2\varepsilon^2 N}.$$

Преобразуем неравенство

$$\frac{|\nu - p|}{\sqrt{p(1-p)}} > 2\varepsilon, \quad (\nu - p)^2 > 4\varepsilon^2(p - p^2).$$

Получили доверительный интервал в форме эллипса.

Случай конечного множества решающих правил

Пусть L – число решающих функций λ .

Имеем

$$\mathbf{P}(\forall \lambda, |\nu_\lambda - p_\lambda| > \varepsilon) \leq L e^{-2\varepsilon^2 N} = e^{\ln L - 2\varepsilon^2 N}.$$

Использовали $\mathbf{P}(\sum A_i) \leq \sum \mathbf{P}(A_i)$.

Получаем

$$\frac{\ln L}{2N} \approx \varepsilon^2.$$

Обозначим $\kappa = \frac{N}{\ln L}$. Тогда $\frac{1}{\kappa} \approx 2\varepsilon^2$.

Получаем соотношение для оценки доверительного интервала

$$\kappa(\nu - p)^2 \approx 2(p - p^2).$$

Доверительный интервал для риска

Доверительный интервал для R зададим в виде $[0, \hat{R}(V)]$, где $\hat{R}(V)$ – оценочная функция или просто оценка (риска). При этом должно выполняться условие:

$$\forall c, P_c(R \leq \hat{R}(V)) \geq \eta,$$

где η – заданная доверительная вероятность.

На практике интервальную оценку будем строить как $\hat{R}(\bar{R}(V))$ – функцию точечной оценки.

Качество интервальной оценки будем характеризовать величиной $E\hat{R}(V)$, которая зависит от c , в виду чего выбор наилучшей оценки становится многокритериальной задачей.

Биномиальное распределение

Пусть p – вероятность «успеха» в схеме Бернулли. Для фиксированного классификатора в роли p будет выступать вероятность ошибки.

Обозначим через $\xi = \frac{N_e}{N}$ – случайную величину, представляющую собой долю ошибочно классифицированных объектов обучающей выборки,

$$B(\gamma, N, p) = P(\xi \leq \gamma) = \sum_{0 \leq i \leq N\gamma} C_N^i p^i (1-p)^{N-i} -$$

кумулятивное биномиальное распределение.

Имеем $P(\xi \leq \gamma) = B(\gamma, N, p)$ – вероятность получить долю ошибочно классифицированных объектов (эмпирический риск) меньше γ .

Если приравнять данную вероятность заданному уровню значимости α , то получим уравнение, связывающее p и γ .

Оценочная функция

Выразив p как функцию γ , получим границу доверительного интервала для вероятности ошибочной классификации.

Пусть $\hat{p}(\gamma)$ – функция, задаваемая уравнением $B(\gamma, N, \hat{p}(\gamma)) = \alpha$. Очевидно, что для любого p выполняется $P(p > \hat{p}(\xi)) \leq \alpha$.

Рассмотрим конечное множество классификаторов Λ , $|\Lambda| = L$. Для каждого $\lambda \in \Lambda$ определена вероятность ошибочной классификации $p(\lambda)$.

Оценки Вапника-Червоненкиса

Обозначим $A(\lambda)$ – событие $p(\lambda) > \hat{p}(\gamma)$.

Имеем

$$\mathbf{P}(p(\lambda(\nu)) > \hat{p}(\gamma)) \leq \mathbf{P}\left(\sum_{\lambda \in \Lambda} A(\lambda)\right) \leq \sum_{\lambda \in \Lambda} \mathbf{P}(A(\lambda)) \leq \alpha L.$$

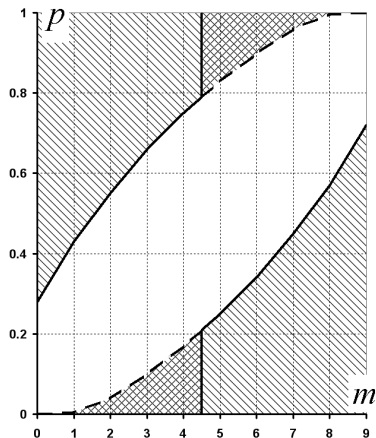
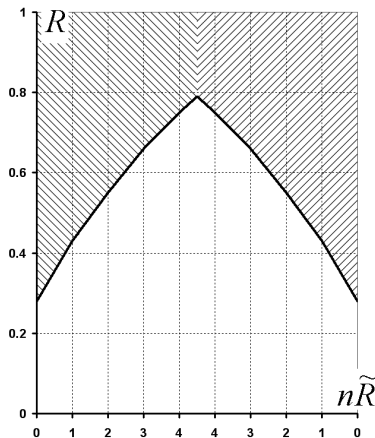
При доверительной вероятности $\eta = 1 - \alpha L$ функция $\hat{p}(\gamma)$ является доверительным интервалом для риска при любом методе $\lambda(\nu)$ выбора решающей функции из Λ .

В оценке присутствуют три неравенства, которые вносят погрешность, объясняемую соответственно

- эффектом «расслоения» классификаторов,
- эффектом «сходства» классификаторов,
- эффектом дискретизации (несущественен).

Пример критического множества

Пусть $L = 2$, $\lambda_2(x) = 1 - \lambda_1(x)$, $p = P(y = 0)$, m – количество объектов $y = 0$ в выборке.



Точные асимптотические оценки

При $N \rightarrow \infty$, $\kappa = \frac{N}{\ln L} = \text{const}$ оценка Вапника-Червоненкиса принимает вид

$$H(\gamma, \hat{p}(\gamma)) = \frac{1}{\kappa},$$

где $H(\gamma, p) = \gamma \ln \frac{\gamma}{p} + (1 - \gamma) \ln \frac{1-\gamma}{1-p}$.

Решение $\hat{p}_\kappa(\gamma)$ данного уравнения оказывается неуллучшаемой (без использования дополнительной информации) асимптотической оценкой риска на основе эмпирического риска.

Пример, доказывающий неуллучшаемость

Пусть дан набор бинарных переменных X_1, \dots, X_L и множество классификаторов $\lambda_1, \dots, \lambda_L$, причем λ_j приписывает объекту класс, номер которого равен значению j -й переменной, то есть $f_{\lambda_j}(x) = x_j$.

Распределение в D задается следующим образом:

$$P(x, y) = P(x | y) P(y); \quad P(x | y) = \prod_{j=1}^L P(x_j | y);$$
$$P(y = 0) = p_0; \quad P(x_j \neq y | y) = p,$$

где p_0 и p — параметры распределения, причем выбор p_0 не имеет значения.

Оценка риска

По построению, риск для любого классификатора λ_j равен p .
Точная вероятность выхода из доверительного интервала:

$$\mathbb{P}(p(\lambda(\nu)) > \hat{p}(\gamma)) = 1 - (1 - B(\gamma, N, p))^L = 1 - \eta.$$

Оказывается, что погрешность оценки union bound (вероятность суммы через сумму вероятностей) для независимых событий в этом контексте несущественна.

Классификация в дискретном пространстве

Пусть $X = \{1, \dots, k\}$. Тогда вероятностная мера $P_c[D]$, $c \in C$, задается набором вероятностей

$$\alpha_j = P(x = j), \quad p_j = P(y = 0 \mid x = j).$$

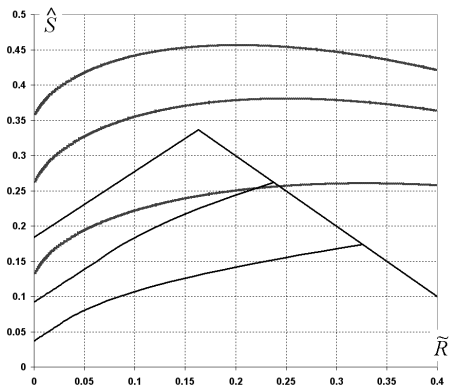
Выборка представляется совокупностью пар

$$V = (v_j \mid j = \overline{1, k}), \quad v_j = (m_j, n_j).$$

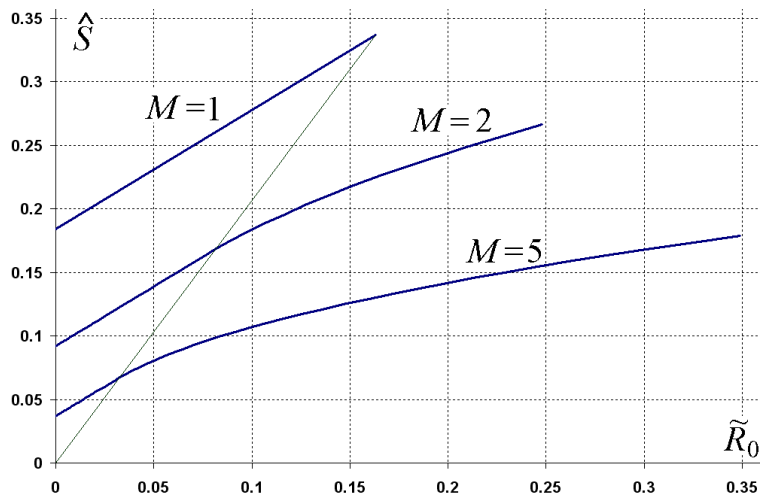
Решающая функция минимизирует эмпирический риск независимо в каждой точке $x \in X$: $f(x) = I(m_j < n_j)$. Рассмотренный случай соответствует гистограммному классификатору.

Точные оценки в дискретном случае

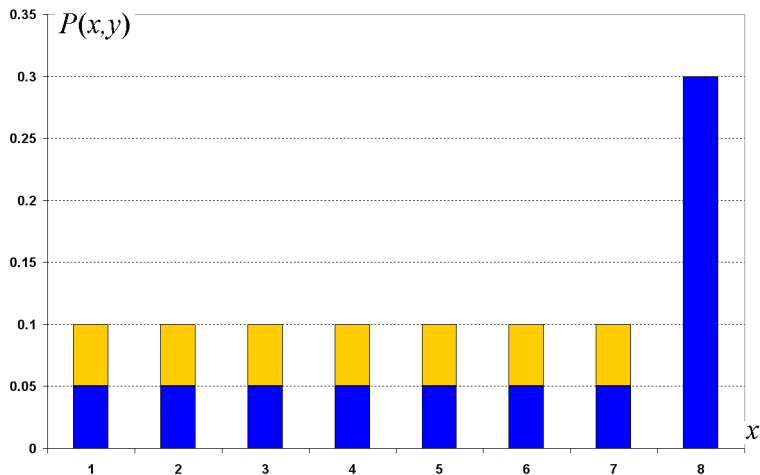
При $N \rightarrow \infty$, $\kappa = \frac{N}{\ln L} = \text{const}$ для дискретного случая найдены точные зависимости для $\hat{S} = \sup_{c \in C} \mathbf{E} R - \mathbf{E} \tilde{R}$.



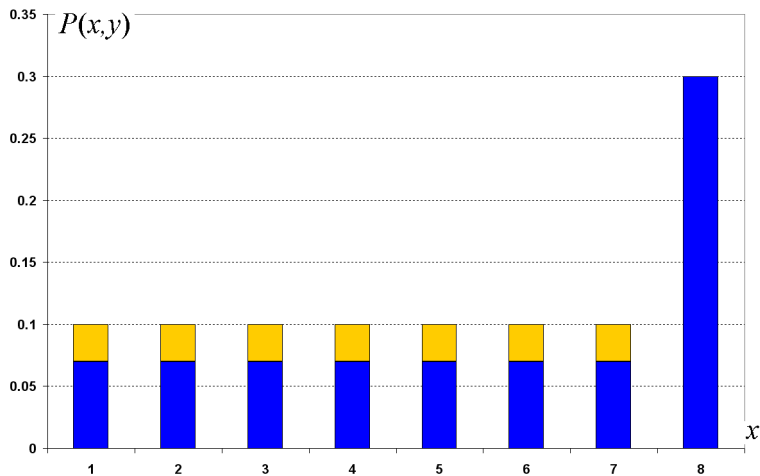
Вид семейства оценок кривых



«Наихудшее» распределение



При малых рисках



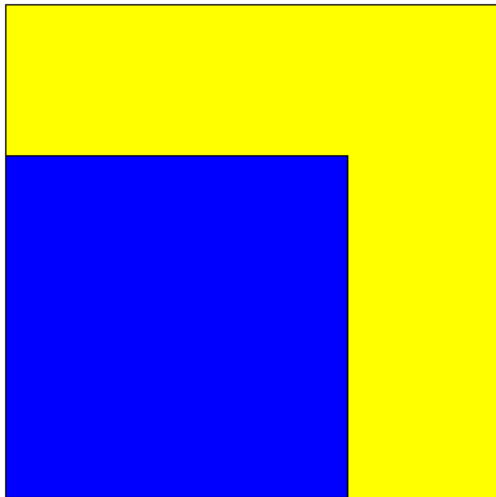
Отличие общего случая

В дискретном пространстве из k точек возможно $L = 2^k$ решающих правил.

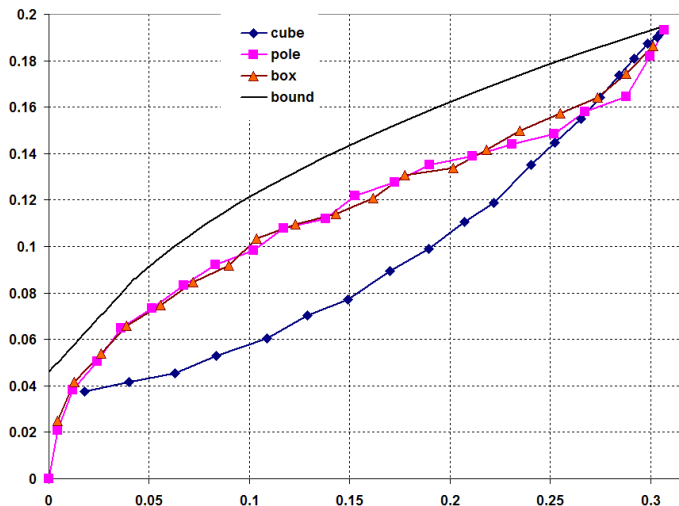
В примере (Langford, 2002), демонстрирующем достижимость оценок Вапника-Червоненкиса, число точек пространства составляет $k = 2^L$.

Гипотеза: для алгебраически замкнутых классов решающих правил асимптотические оценки риска не превосходят оценки для дискретного случая.

Параметрические семейства модельных распределений



Результаты для деревьев решений



Замечания

- оценки Вапника–Червоненкиса существенно не улучшаемы без использования более полной информации,
- максимальное смещение эмпирического риска достигается на двух семействах распределений и аппроксимируется простыми выражениями,
- оценочная кривая единственна с точностью до подобия,
- оценивание параметра сложности может быть произведено моделированием на «нулевой» стратегии,
- «наихудшее» распределение неоднородно по степени «перемешанности» классов,
- найденные варианты «наихудших» распределений могут использоваться при построении эмпирических доверительных интервалов.